Capstone project: Rating my professor data analysis

Group Name: [Group 77]

Authors Contribution:

Most problem-solving decisions were made collaboratively. The contributions were divided as follows:

Xu, Haoying(focused on the first six parts utilizing statistical expertise):

Data cleaning, normalizing label counts, data weighting, data visualization (with support from ChatGPT), and selecting test statistics for the first six questions.

Report organization and debugging.

Wu, Bokai (focused on the remaining four parts leveraging computer science knowledge):

Data cleaning, data scaling before fitting, model selection, constructing pipelines for model implementation, and data visualization for the last four questions.

Code instruction and implementation, and report review.

Data Preprocessing and Cleaning:

In this project, data preprocessing was handled as follows:

Data Cleaning:

For questions (1) to (6), which aim at comparisons in terms of gender:

We filtered out professors whose gender couldn't be determined by deleting rows where columns 7 and 8 contained the patterns '00' or '11' in the *rmpCapStoneNum.csv* dataset. We prioritized the quality and reliability of the data over the sample size. As a result, we retained 52,089 data points with no missing values for the analysis of the first six questions.

For questions (7) and (10):

Based on the 52,089 data points obtained above, we applied row-wise element cleaning to filter out professors whose attributes were empty in the *rmpCapStoneNum.csv* dataset. We retained 8,849 data points, which we considered an acceptable sample size. Additionally, we deleted column 8 to prevent the dummy variable problem.

For questions (8) and (9):

Since the *rmpCapStoneTags.csv* dataset had no missing values, we only filtered out professors whose average ratings or difficulty scores were null.

Data Transformation:

For questions (1) to (6):

It is known that sample averages based on more ratings are more meaningful and converge to the true mean according to the law of large numbers. Therefore, when comparing the average rating/difficulty between two groups, we transformed the average rating/difficulty data into weighted data by replicating (or, in other words, 'attaching importance to') each data point according to the number of ratings before performing statistics inference. This approach aimed to increase the influence of averages from larger samples while maintaining the representativeness of those from smaller samples.

For the raw tag data in *rmpCapStoneTags.csv*:

Before performing any statistical inference or machine learning processes on the raw tag data in *rmpCapStoneTags.csv*, we decided to normalize all the tag data by transforming them into tag rates (raw number of tags / number of ratings). Since professors with more ratings tend to receive more tags, normalizing by the number of ratings helped mitigate this confounding effect. Tag rates provided a fairer metric for evaluating a professor's ability to acquire tags, making it a more meaningful measure for comparison.

Data scaling: Since regularization is sensitive to the magnitude and variance of the features, we scaled the data using StandardScaler() from sklearn.preprocessing before fitting the model. This ensures that all features are on the same scale, which helps prevent larger-magnitude features from disproportionately influencing the regularization process.

This report provided a comprehensive analysis of professors, analyzing variability as well as factors that influenced, among other things, student evaluations of professors. The dataset included numeric ratings, qualitative data, and labels. We addressed ten key questions to understand gender bias, variability, and factors that influence students' ratings of professors.

Question 1: Evidence of Pro-Male Gender Bias

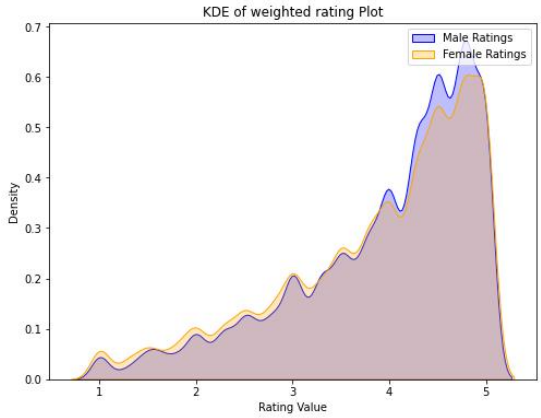
To investigate whether male professors enjoy a boost in ratings due to gender bias, we performed a one-tail Mann-Whitney U test to compare median, as the data on male and female ratings showed non-normal distribution (skewness check). The preproccessed data was split into two parts according to the gender, and then the average rating data of two groups was compared after weighing. The results showed a U-statistic (9610656830)and a p-value (5.720451156012289e-52) below 0.005, suggesting that male professors have statistically higher ratings, implying a potential pro-male bias, so we accept H1. The KDE plot visually supports this difference, as the distribution for males is slightly shifted toward higher values compared to females. However, such effect size might be negligible according the following cliff's Delta.

Effect Size: Cliff's Delta = (0.03346740772404044, 'Negligible effect')

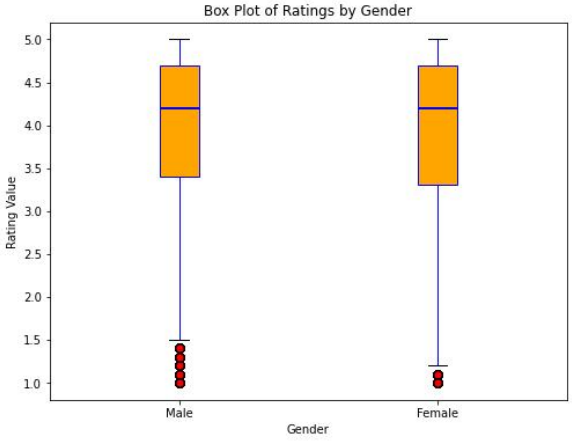
Hypothesis

H0: The (median of)average rating between male and female professors was same

H1: The (median of)average rating of male professor was higher than that of female professor



Question 2: Gender Differences in Rating Dispersion



The spread in ratings was compared using the **Brown-Forsythe test** for equality of variances, with the **median** as the center due to the skewness of the data. The test yielded a significant result ($p < 0.005$), indicating a difference in the variance of ratings between male and female professors. However, the **median absolute deviation (MAD)** for both groups was the same (0.6), suggesting that the dispersion difference between the two groups was small. As observed in the boxplot, this significant result may be attributed to the influence of low-rating outliers, as the Brown-Forsythe test is relatively sensitive to outliers. In conclusion, the difference in dispersion between the two groups, excluding the outliers,

appears to be small, aligning with the interpretation of the left boxplot.

Hypothesis

H0: The deviations from the median were equal for male and female professor average rating

H1: The deviations from the median were not equal for male and female professor average rating

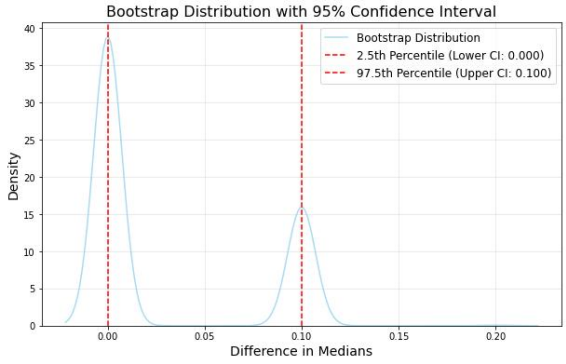
Statistic: Brown-Forsythe: B = 519.0518145040764, p = 8.706432723919317e-115.

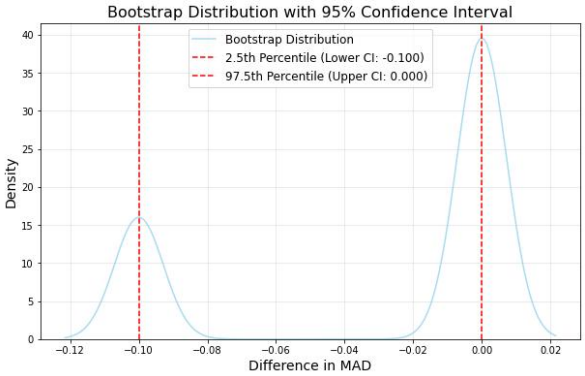
MAD: for both: 0.6

Question 3: Size of Gender Bias Effects ("Diff = Man - Woman")

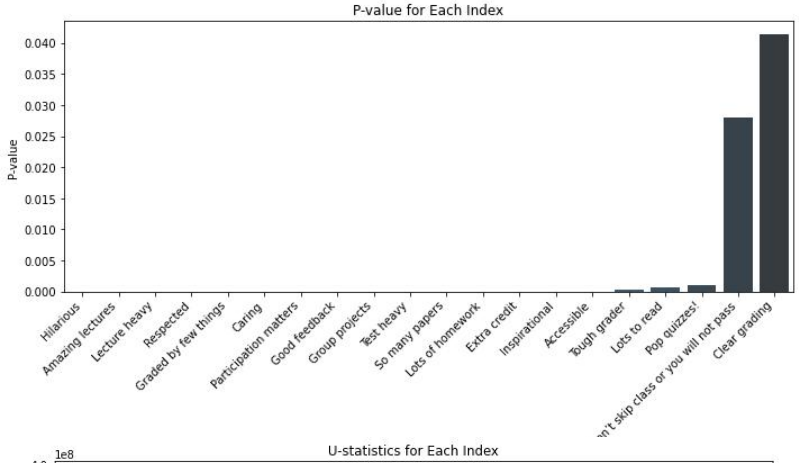
To compute the specific size of the effects in terms of both mean and spread, bootstrap sampling (with 10000 iterations) was used to generate confidence intervals since the distributions of both groups were skewed. The 95% CI for the difference in median ratings was found to be [0, 0.1], while the CI for the difference in MAD (Median Absolute Deviation) was [-0.1, 0], as were shown in the following graph (Bootstrap distribution is estimated using guassian KDE).

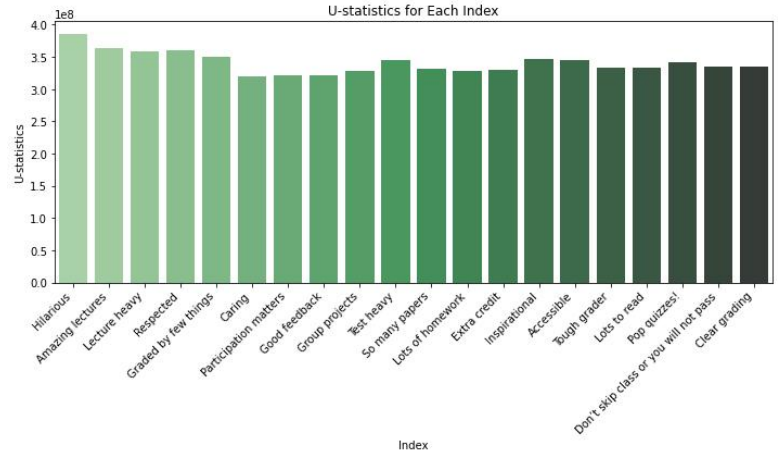
Method: During 10000 iterations, resample each group independently from the weighted data of each group and compute the difference in median and MAD for two groups of resampled data. Then 95% median fell in [0,0.1], and 95% MAD fell in [-0.1,0]



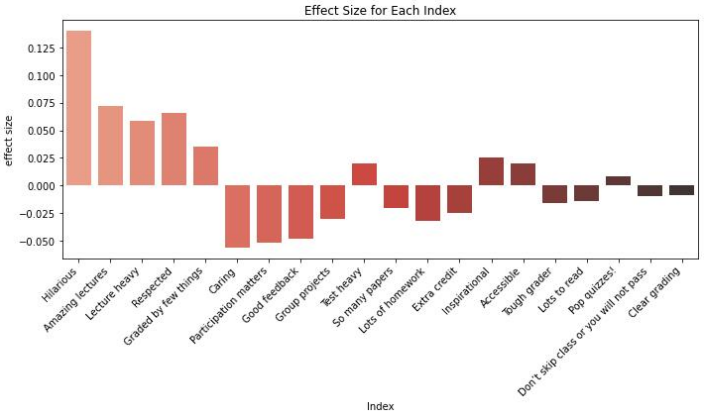


Question 4: Gender Differences in Tag rate





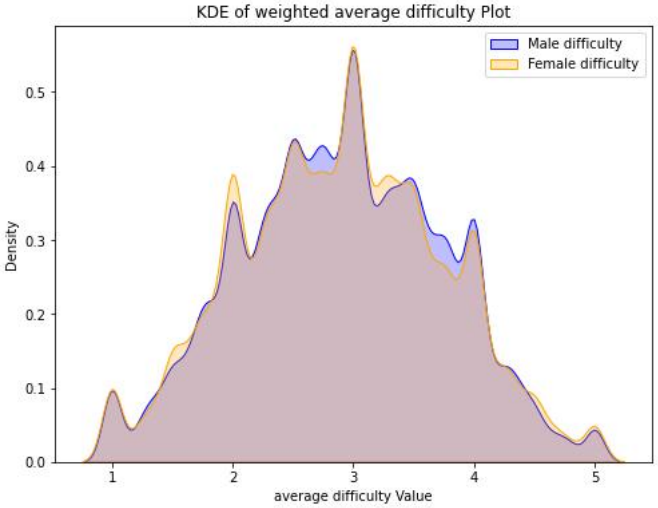
We normalized the tags by dividing the number of times a tag was awarded by the number of ratings. Then split the tags data into two groups according to the gender. For each of the 20 tags, a statistical test was conducted to identify any significant gender differences. Since all the distributions of the tag rates were identified as skewed, the U test was conducted for all comparison. The first graph on the left shows that out of the 20 tags, 18 were found to be significantly different ($p < 0.005$) in terms of the tag rate. The tags with the least significance were "Clear grading", "Don't skip class," and "Pop quizzes.", and the tags with the most significance were "Hilarious", "Amazing lectures" and "Lecture heavy". However, the effect size (cliff's delta) indicated that all differences were at the level of "negligible". We



concluded that there existed gender difference in those 18 tags but the difference was actually small, so these attributes are largely perceived similarly across genders. Since multiple tests were done in this question, the overall significance level(false positive rate) was 0.09 (0.005*18).

For each test from 20 comparisons:
H0: The (median of)average rating between male and female professors were same
H1: The (median of)average rating between male and female professors were different

Question 5: Gender Differences in Difficulty Ratings
As in Q1, we split the average difficulty data into two sets and transformed them into weighted data. Since the distributions of both groups were approximately symmetric and normal, **Welch's t-test** was applied to compare the difference in means, supported by the significant result of the **F-test** ($p < 0.005$). The t-test also yielded a significant p-value ($0.00013 < 0.005$), leading to the conclusion that there is a gender difference in difficulty ratings. However, the effect size measured by **Cohen's d** was 0.0146, indicating a negligible difference in practical terms. Additionally, a **Kolmogorov-Smirnov (KS) test** was conducted to identify differences in the shape of the distributions. The KS test returned a highly significant p-value ($1.28e\text{-}10 < 0.005$), suggesting a difference in the overall shape of the distributions. Despite the statistical significance, visual inspection of the distributions (shown in the graphs below) revealed substantial overlap between the two groups. This suggests that the flattened sample sizes after weighting might have enhanced the sensitivity of the tests to detect even small differences.
F Test: f=1.0270109249167934, p=9.242280392118829e-07
H0: The variances of average difficulty between two groups were same
H1:The variances of average difficulty between two groups were different
t Test: 3.811856937870696, p = 0.00013795816788894196
H0: The(mean of) average difficulties of two groups were same
H1:The(mean of) average difficulties of two groups were different
The cohen's d: 0.014642104720626709
KS Statistic: 0.013146932181391735, p = 1.284383790005226e-10
H0: The distributions of average difficulty of two groups were same
H1: The distributions of average difficulty of two groups were different



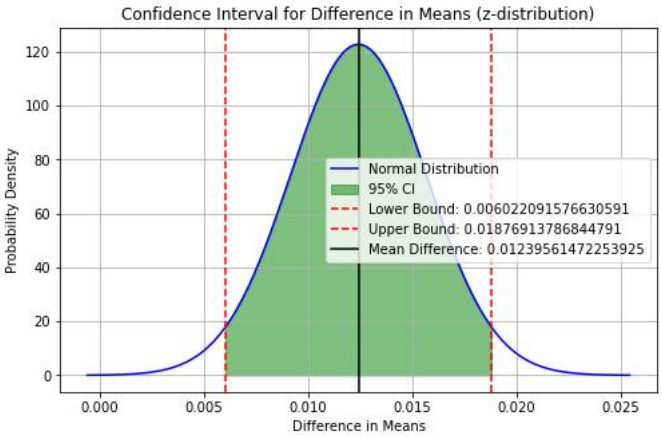
Question 6: Size of Difficulty Rating Effect ("Diff = Man - Woman)

A **95% confidence interval** for the difference in mean difficulty ratings was calculated using the **Central Limit Theorem**, given the large sample sizes for both groups. The confidence interval ranged from 0.006 to 0.018, indicating that any potential difference is small and likely negligible.

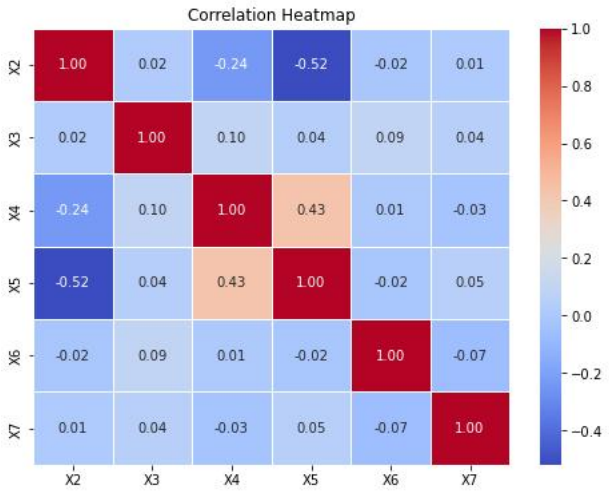
Difference in Means: 0.01239561472253925

95% Confidence Interval:

[0.006022091576630591, 0.01876913786844791]



Question 7: Regression Model for Average Rating



A **Lasso linear regression** model was constructed to predict the average rating using all numerical predictors. First, the correlation matrix (shown on the left) revealed weak collinearity, suggesting that multicollinearity did not need to be addressed. Next, similar to our previous approach, the loss function was weighted according to the number of ratings. The dataset was randomly split into training and test sets with an 80:20 ratio, and cross-validation was used to identify the best hyperparameters that minimize the average mean squared error (MSE) on the validation set. The results are presented below.

The Best lambda for Lasso is : 0.012120900900900903

Lasso Model Equation:

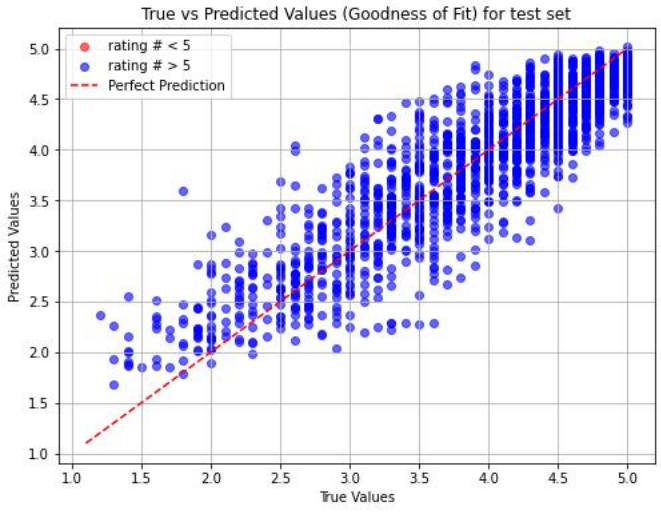
y = 2.5205 + (-0.1964) * X2 + (0.0001) * X3 + (0.1957) * X4 + (0.0249) * X5

RMSE for trainning set: 0.36624208553072723

R² for the trainning set: 0.8070144397476658

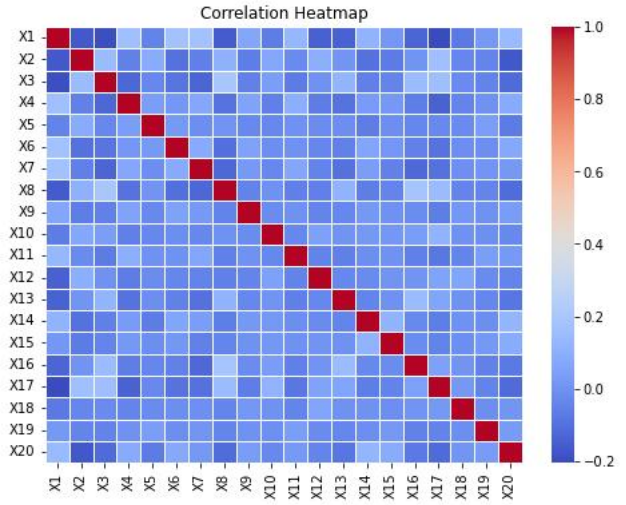
RMSE for the test set: 0.36090056817697785

R² for the test set: 0.8108616709643762



The Lasso linear regression model performed well, explaining 81% of the variance in average ratings and achieving an RMSE of 0.36 for both the training and test sets, indicating predictions were off by only 0.36 units on average. Using the optimal λ=0.0121, the model effectively filtered out irrelevant predictors like gender(X7) and the number of online ratings(X6), while also identifying difficulty(X2) and Pepper status(X4) as the strongest predictors, with "take class again proportion" (X5) moderately contributing to the ratings. The results highlight the significance of course difficulty and recognition in shaping student evaluations, while also demonstrating the model's ability to simplify feature selection and focus on the most impactful variables.

Question 8: Regression Model for Tag rates Predicting Ratings



Another Lasso regression model was built using the normalized tags as predictors. First, the weak multicollinearity needn't to be considered as illustrated by the map on the left. The principle in terms of data splitting, fitting and hyperparameter tuning was same as that in Q7. Then the results are presented below:

The Best lambda for Lasso is : 0.013321981981981984

Lasso Model Equation:

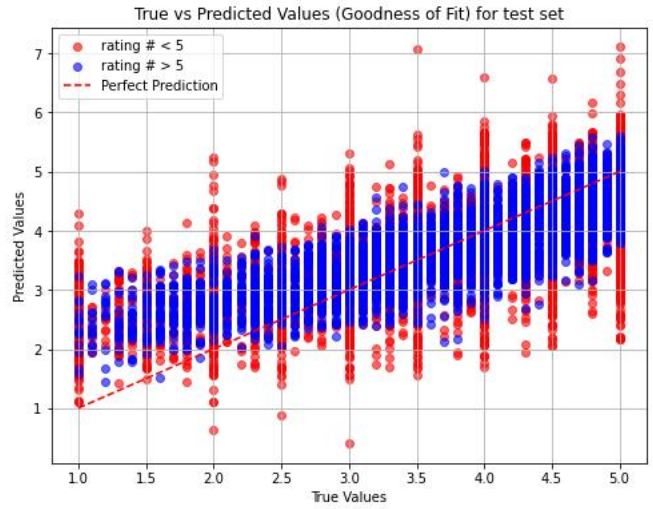
y = 3.2258 + (-1.0435) * X1 + (0.8709) * X2 + (0.8263) * X3 + (-0.0173) * X4 + (0.2234) * X5 + (0.0313) * X6 + (-0.1338) * X7 + (0.5057) * X8 + (0.3655) * X10 + (-0.2896) * X11 + (0.6907) * X12 + (0.6299) * X13 + (-0.5717) * X14 + (-0.2954) * X15 + (1.0257) * X16 + (0.7358) * X17 + (0.3774) * X18 + (-0.0425) * X19 + (-0.4930) * X20

RMSE for trainning set: 0.7949674376157926

R² for the trainning set: 0.5005075261435753

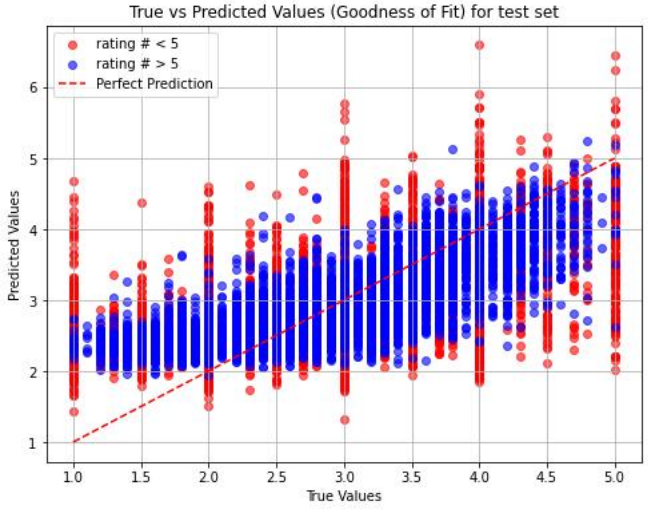
RMSE for the test set: 0.7968041667178645

R² for the test set: 0.5072386903827557



The Lasso regression model trained on the data from rmpCapstoneTags.csv performed modestly, explaining only 50% of the variance and achieving an RMSE of 0.8, which is considerable given the rating scale of 1 to 5. By selecting an optimal λ, the least relevant feature, X9(Pop quizzes!), was removed, while X1(Tough grader) and X16(Amazing lectures) emerged as the most predictive features with coefficients exceeding 1. When evaluated on the test set used in the previous model, this model showed improved performance (R^2=0.73, RMSE = 0.43), though it still lagged behind the numerical predictor model, so the model using the numerical predictors is better. In addition, this improvement likely stems from the numerical predictor model's exclusion of data from participants with fewer ratings and the weighted loss function where the model was based. As it could be seen from the goodness of fit graph for the test set of this model, the participants with few ratings introduced higher prediction errors(marked by the red points) and hence contributed to more unexplained variance. The results suggest that the tag-based model performs better for data with more ratings, where it can more accurately predict outcomes.

Question 9: Regression Model for Tag rates Predicting Difficulty



The dataset to build this model was same as that for the previous one except that the response was replaced by the average difficulty. Apply the same principle of model construction as the previous one. We got the following results:

The Best lambda for Lasso is : 0.00941846846846847

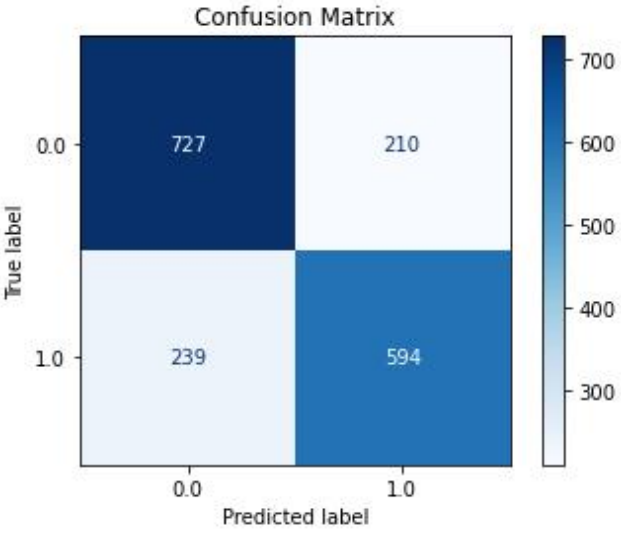
Lasso Model Equation:

y = 2.6193 + (1.5693) * X1 + (-0.0717) * X2 + (-0.1270) * X3 + (0.3492) * X4 + (-0.0845) * X5 + (0.3351) * X6 + (0.3422) * X7 + (-0.0591) * X8 + (0.0776) * X9 + (0.7380) * X10 + (0.0722) * X11 + (-0.4293) * X12 + (-

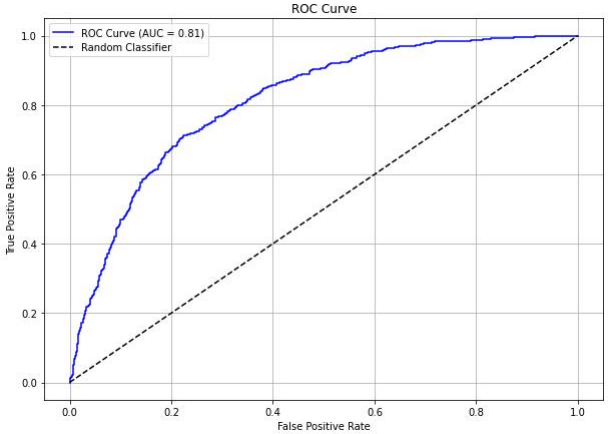
0.3470) * X13 + (1.1726) * X14 + (-0.1384) * X15 + (-0.2180) * X17 + (-0.3190) * X18 + (-0.0282) * X19 + (0.1416) * X20
RMSE for trainning set: 0.8110954656436652
R² for the trainning set: 0.32835272415566363
RMSE for the test set: 0.8195948561461033
R² for the test set: 0.32348625498610883

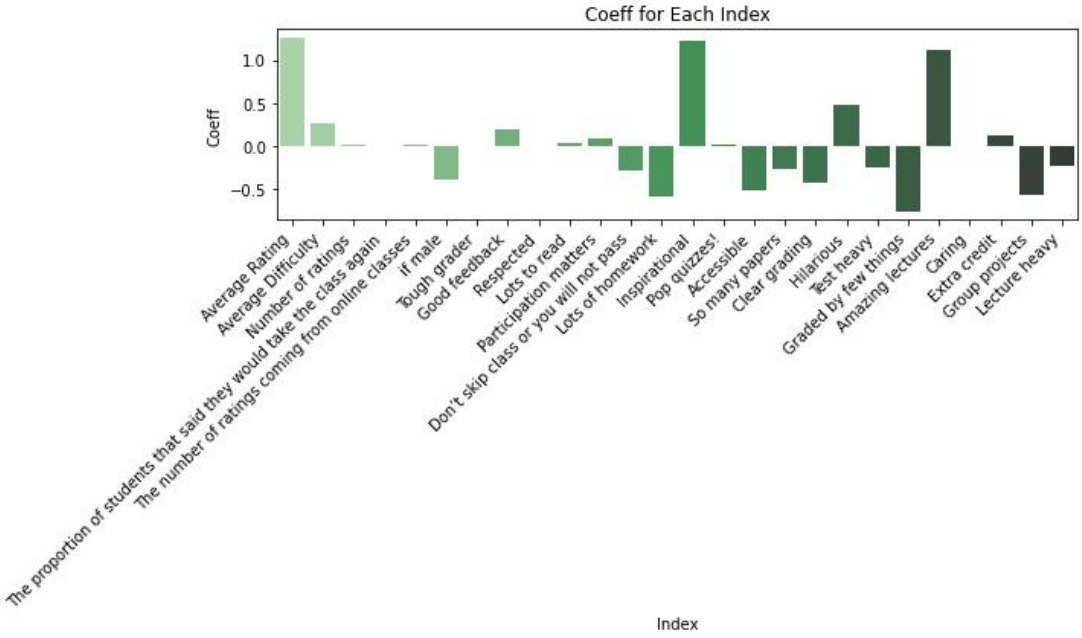
This model only explained 32% of the variance in average difficulties for both set, and its predictions were off 0.8, which was considerable given the difficulty scale of 1 to 5. By selecting an optimal λ, the least relevant feature, X16(Amazing lectures), was removed, while X1(Tough grader) and X14(Test heavy) emerged as the most predictive features with coefficients exceeding 1. Like the previous model, this model struggled to predict those professors with few ratings(marked by the red points), but this model still performed modestly with 0.57 R^2 and 0.48 RMSE on the participants from the test set of Q7. Therefore, there might exists nonlinear relationship between difficulty and tag rates, or that more predictors are needed. However, besides prediction, the model still provides useful insights into the individual impact of each feature on difficulty ratings, which can guide further analysis.

Question 10: Classification Model for "Pepper" Award



A logistic regression model with Lasso regularization was used to predict whether a professor received a "pepper." There were 4166 “1” and 4683 “0”, which demonstrated a slight imbalance of class. To preserve the distribution of the whole dataset and minimize bias caused by random sampling, we applied the **stratified sampling method** during the random splitting process. Then, we set the *class_weight parameter* to *'balanced'* in the logistic regression model provided by scikit-learn so that misclassifying the minority class is penalized more heavily during training. After the search for the best hyperparameter that maximized the AUROC during the cross-validation proccess, the model with the optimal hyperparameter had an AUROC score of 0.81, indicating reasonable discriminatory power. A confusion matrix was also plotted to evaluate the model's predictions, and a threshold was chosen to maximize the true positive rate minus the false positive rate. The numerical results are as follow:
Best hyperparameters: {'C': 0.0872972972972973}
Test ROC-AUC Score: 0.8135399303798359
Best Threshold: 0.57
Intercept: -6.55416384
Based on the threshold we chose, we computed the confusion matrix with respect to the test data. It can be inferred from the matrix that most metrics(such as accuracy,precision and recall) to evaluate this model are over 70%, so we concluded that this model performed well on the classification task on “pepper” identification. As is seen from the graph below, such irrelevant features as “tough grader”, “Respected”, “Caring” and “The proportion of students that said they would take the class again” were eliminated through the regularization for the purpose of the simplification of the model, and the most strongly predictive features are “Average Rating”, “Inspirational” and “Amazing lectures”. (ifmale(gender) corresponds to the 7th column in *rmpCapStoneNum.csv)*




Coeff for Each Index
Coeff
1.0
0.5
0.0
−0.5
Average Rating
Average Difficulty
Number of ratings
The proportion of students that said they would take the class again
The number of ratings coming from online classes
If male
Tough grader
Good feedback
Respected
Lots to read
Participation matters
Don't skip class or you will not pass
Lots of homework
Inspirational
Pop quizzes!
Accessible
So many papers
Clear grading
Hilarious
Test heavy
Graded by few things
Amazing lectures
Caring
Extra credit
Group projects
Lecture heavy
Index